Submission by Reed Elsevier in response to Irish Copyright Review Committee (CRC) consultation paper:
Response by: Peter Carroll (scibella.com)
Date: 29th June 2012

I am involved in a project that seeks a transformative re-use of information about current scientific research projects to provide a Current Research Information Service (CRIS) for the UK and Ireland. I have previously submitted to the CRC as Peter Carroll
[1:http://www.djei.ie/science/ipr/carroll_peter.pdf]

The CRC have published the submission from Reed Elsevier in response to their consultation paper as[2] http://www.djei.ie/science/ipr/Reed_Elsevier.pdf . They confine their detailed comments and evidence to the concept of a new exception for text and data mining discussed in section 9.8 of the Consultation paper.

In my response I italicise excerpts from Reed Elsevier's submission. My response is in normal text.

Over the last decade Reed Elsevier has invested some £2.02bn to combine technology with authoritative content to deliver enhanced functionality to the user. [2,p.2]

This seems an impressively large sum. Presumably most of this investment was in IT. However, it covers the expenditure of the whole Reed Elsevier publishing operations not just their Scientific Technical & Medical (STM) publishing. Total turnover reported by Reed Elsevier for the decade 2001-2010 is £51.9 bn. [Source annual reports: http://www.reedelsevier.com/investorcentre/Pages/results-centre.aspx]. Expenditure of £2.02 bn is about 3.8% of turnover- comparable to the 3.4% expended on IT by large US companies in 2007 [source: http://www.metrics2.com/blog/2006/06/26/average_company_spends_34_of_revenue_on_it.html]

What would be more relevant would be their investment in the STM areas directly affected by a text/data mining exception. Particularly investment in the ConSyn system introduced in 2010 to *"offer the facility to digitally transfer bulk copies of the required content to the user via a secure delivery mechanism which we call ConSyn" [2 p.3]*

3. Where is the market failure?

We would suggest that far from there being a market failure there is extensive evidence that publishers are creating opportunities for researchers to text and data mine journal content. We quote our own business offerings in the following section as examples.

Against this is the UK report by JISC into text/data mining [3 source: http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx] which looked into the issue of market failure and concluded:

"There appears to be strong evidence for market failure against three of the four indicators for market failure in the 'Green Book'. Strong evidence for market failure can justify government intervention and would tend to support the case for the Hargreaves recommended copyright exception on the grounds of improving economic efficiency." [3, sec. 5.3]

Also there is direct evidence from researchers, trying to use content mining, of problems with STM publishers see: http://www.richardpoynder.co.uk/Content_Mining.pdf [4]for examples.

In addition, we submit that significant progress is being made on cross-corpora mining (i.e. across different publisher collections). The STM publishing industry has developed a model licence that will simplify the contracting process,and industry discussions about how to achieve the necessary technical standardisations are now underway.

The model licence from STM can be found at http://www.stm-assoc.org/2012_03_15_Sample_Licence_Text_Data_Mining.pdf

I am not a lawyer but this model licence appears to reserve many rights to STM publishers and grant very few to users. The *cross corpora mining* presumably refers to the Linked Content Coalition project http://www.linkedcontentcoalition.org/Home_Page.html [5] set up by the European Publishers Council. Progress to date can be gauged by the report of the first Plenary session held on 20 June 2012. http://www.linkedcontentcoalition.org/uploads/1206120_plenary.pdf

We note that the Consultation paper references the UK Hargreaves report and the UK Government's response to Hargreaves when discussing a possible exception for text and data mining. As noted above, we believe changes to IP policy should be guided by evidence and would draw the CRC's attention to the fact that after months of work, both by Hargreaves and the UK's Intellectual Property Office, the UK consultation document and impact assessment did not identify any monetary benefits in favour of the exception, or accounts of costs.

Estimates of the economic benefit to the UK research community of text/data mining are set out in the JISC report [3: section 5 Economic analysis of the value and benefits of text mining in UKFHE].

"If text mining enabled just a 2% increase in productivity – corresponding to only 45 minutes per academic per working week[37] (and looking at CIBER's analysis of the impact of eJournals [69], this is very much an underestimate), this would imply over 4.7 million working hours and additional productivity worth between £123.5m and £156.8m in working time per year."

Publishers are enabling Text and Data Mining

As part of Elsevier's commitment to enable access to support innovation, in late 2010 we opened up our core database ScienceDirect (10 million full text articles and book chapters) and Scopus, the world's biggest database of academic abstracts and citations (19,000 titles from more than 5,000 international publishers), for the building of applications by third party developers.

This means that it is possible for anyone from an Irish university or Irish technology business to use our content to develop text and data mining tools. There is no charge and the tools are mounted on our platforms for the collective benefit of our researcher users. The IP in the developed applications remains with the developer, and is non-exclusive, meaning that they may also offer it to other publishers' platforms and users. We provide access to the content via application programme interfaces (APIs) after validating the credentials of the developers to ensure content security. By having the applications on our platform we can utilise our entitlement system to enable users to mine subscribed or Open Access content.

Professor Peter Murray Rust from the University of Cambridge has a different take on STM publishers offering access via APIs on their systems.

“P M-R: I don't want to use Elsevier's API. That means 100 APIs for me to learn — one per publisher.

In fact, I only need a single API — a DOI resolver. I may wish to systematically mine a single publisher — in which case I use a list of their DOIs, or I may want to follow links — that's exactly the same process. Yes I need an API per publisher but I and others are hacking this and it's a one-off. So a publisher API makes it worse.

The only conceivable argument for publishers to insist on the use of APIs is server load. Elsevier publishes 250,000 papers each year, has an archive of 7 million publications, and has 240 million downloads. That's a soluble problem.” [4, p14]

Whilst a publisher provided API may be sufficient for many researchers there are many who need to analyse content with their own software. As Cameron Neylon puts it:

“Because the mining I want to do and the mining that Peter Murray-Rust wants to do will be different, and what I will want to do tomorrow is different to what I want to do today. This kind of personalised mining is going to be the accepted norm of handling information online very soon and will be at the very centre of how we discover the information we need”

[6 source: http://cameronneylon.net/blog/they-just-dont-get-it/]

Impact on platform stability
We remain concerned that the exception would lead to aggressive crawling by multiple parties of our core publishing platform, with knock on effects of significantly reduced user experience. Currently we only allow Google to crawl full text articles.

Cameron Neylon again:

“The technology exists today to make this kind of mass distributed text mining trivial. Publishers could push content to bit torrent servers and then publish regular deltas to notify users of new content. The infrastructure for this already exists. There is no infrastructure investment required. The problems that publishers raise of their servers not coping is one that they have created for themselves. The catch is that distributed systems can’t be controlled from the centre and giving up control requires a different business model. But this is also an opportunity. The publishers also save money if they give up control – no more need for six people to sit in on each of hundreds of thousands of meetings. I often wonder how much lower subscriptions would be if they didn’t need to cover the cost of access control, sales, and legal teams.” [6]

What is research?
The Consultation paper does not try to define text and data mining. The UK consultation describes text and data mining as an automated text and data analysis for patterns, trends and “other useful information”. That seems to describe a search engine query. What is a search if not an "automated analytical technique" (i.e. an algorithm) that is used for picking out "useful information" or "patterns" from digitised data and text? All searches are a form of research.

A proposed “working” definition for text and data mining can be found at the STM association, of which Reed Elsevier are members, web site:

http://www.stm-assoc.org/2012_03_15_STM_Summary_Statement_Text_Data_Mining_final.pdf

"A computational process whereby text or datasets are crawled by software that recognises entities, relationships and actions."

As is made clear in the STM statement this definition aims to distinguish text and data mining from other activies such as Information Retrieval and Information Extraction.

Disincentive to supply to the Republic of Ireland
The risk of damage to the core business of journal publishing posed by the exception, as outlined above, could make publishers consider whether it is worthwhile continuing to offer their publications in Ireland. As Irish law will apply regardless of the origin of the work, it is foreseeable that the owners of foreign copyrights could set up firewalls preventing access by users located in the Republic. E-sales in Ireland only represent approximately 0.2-0.3% of total STM global e-revenue. To avoid the significant downside risks, rights holders may end up choosing to forgo the Irish market, thereby depriving Irish users of access to valuable content.

What can I say. Put more succinctly - "Drop the exception or the scientist gets it". In areas other than copyright reform the phrases censorship and "collective punishment" might come to mind.

References:

[1] http://www.djei.ie/science/ipr/carroll_peter.pdf

[2] http://www.djei.ie/science/ipr/Reed_Elsevier.pdf

[3] http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

[4] http://www.richardpoynder.co.uk/Content_Mining.pdf

[5] http://www.linkedcontentcoalition.org/Home_Page.html

[6] http://cameronneylon.net/blog/they-just-dont-get-it/